Discrete and Continuous Compositions

J. Bacon-Shone

Social Sciences Research Centre, The University of Hong Kong, Hong Kong

johnbs@hku.hk

Abstract

This paper examines a dataset which is modeled well by the Poisson-Log Normal process and by this process mixed with Log Normal data, which are both turned into compositions. This generates compositional data that has zeros without any need for conditional models or assuming that there is missing or censored data that needs adjustment. It also enables us to model dependence on covariates and within the composition.

Key words: discrete composition, Poisson distribution, structural zeros

1 Introduction

Many things that we measure and treat as if they are continuous are really discrete count data, even if only at molecular extremes. When we form compositions from count data. the underlying discrete nature of the data may be hidden, except for the occurrence of zeros.

Another perspective is that there are times when compositions may clearly not be generated directly from logistic normal distributions or indirectly by applying the compositional process to multivariate log normal distributions, but by applying the compositional process onto other multivariate distributions on R^{d+}. If we know what those distributions are, it is obvious that we should use that information, although logistic normal distributions may provide a useful approximation in some situations.

In this paper, we examine two alternative ways of generating compositions and how close those distributions are to the logistic normal and the consequences in terms of zero components. These alternatives are:

i. Poisson-Log Normal distribution of Aitchison Ho (1989) generating counts and then forming a composition

ii. The same count data mixed with log normal and then forming a composition

2 Dataset

We revisit the goilbird dataset of Aitchison (2003) presented at CoDaWork 03. Fortunately, we discovered that some additional data was collected at the same time for the 60 goilbirds. In addition to the time budget data previously noted, we have data relating to the feeding element. As any visitor to Scotland will warn you, there are lots of little insects at certain times of year, which humans dislike but the goilbirds happily devour. Firstly, we have records of how many insects of three different types (creepies, crawlies and flies) each goilbird caught during the feeding period. It was not feasible to weigh the insects directly, but we can estimate the weights caught per bird based on the size measured from photographs. The data can be found in the Appendix.

2. Modelling the counts

Figure 1 Bivariate Fit of Total Count By Feed

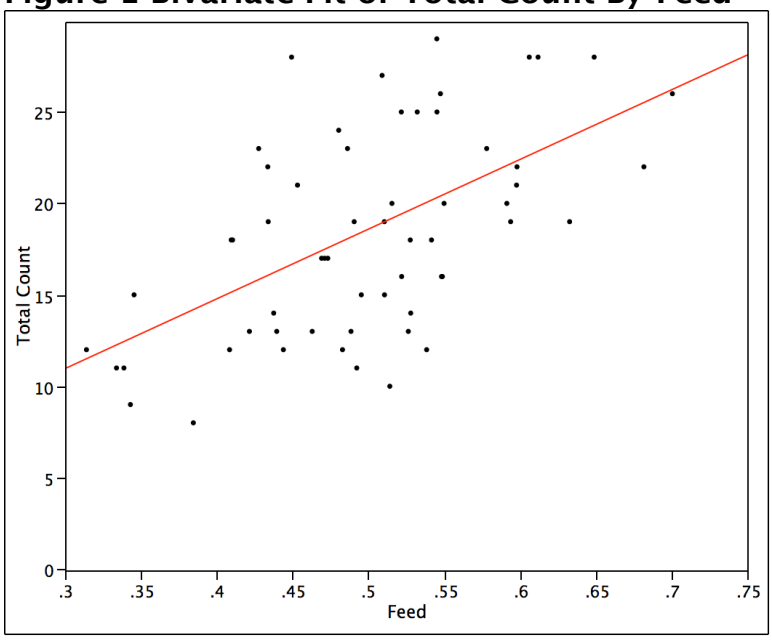
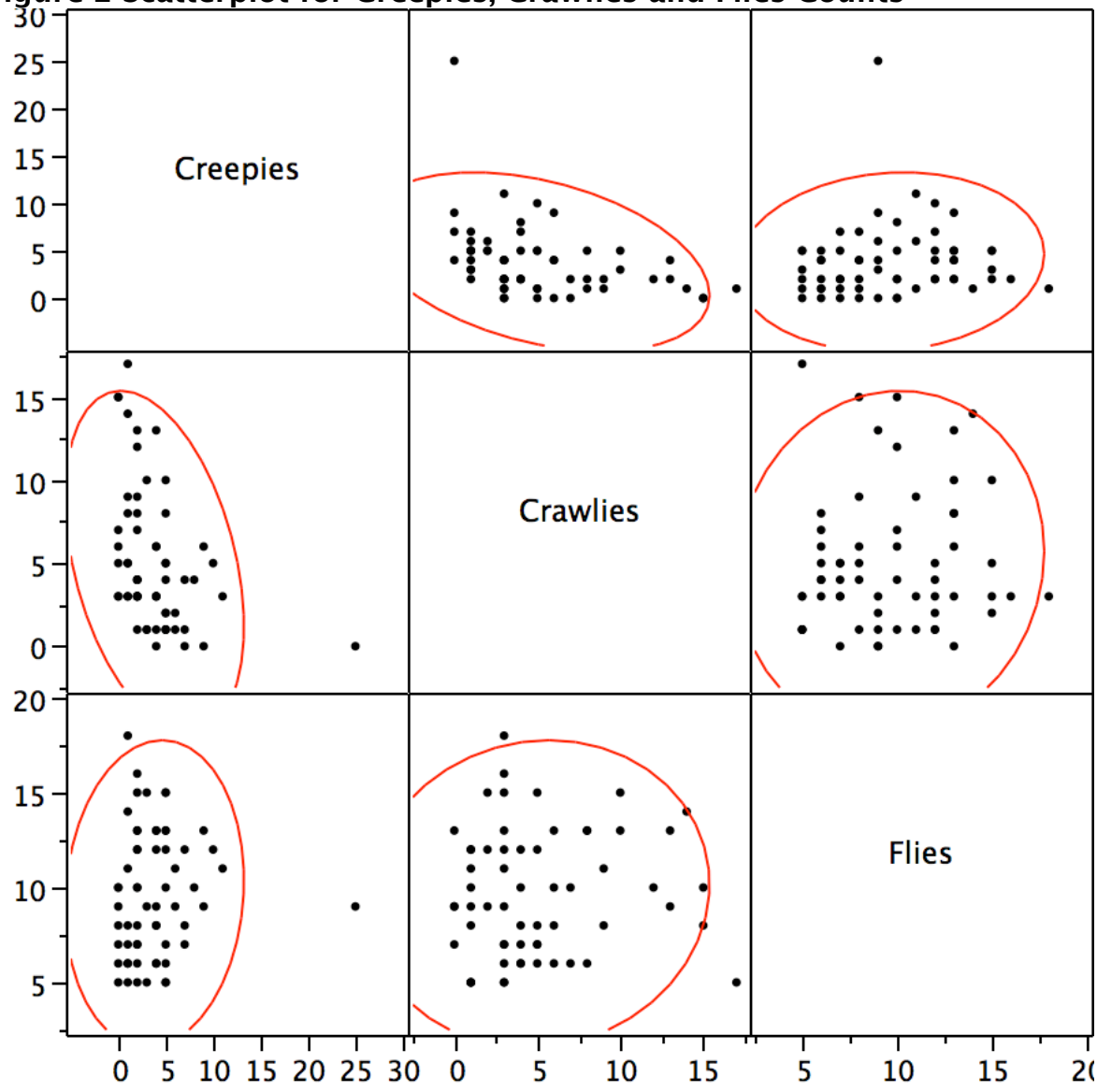


Figure 1 shows a plot of the total number of insects caught by each bird against the proportion of time spent feeding, which shows a clear linear relationship. A similar relationship appears for the counts of the individual species. This suggests that the counts follow a Poisson process with mean proportional to the time spent feeding and indeed the fit of such a model is good.

Figure 2 Scatterplot for Creepies, Crawlies and Flies Counts



However, scatterplots of the counts in Figure 2 make it clear that the counts of the different species are not independent. The plot suggests negative correlation of the relative creepie and crawlie numbers, perhaps because of competition or choice between them by the birds, while the fly numbers may be independent. Note that because of the common effect of feeding time on all counts, there should be some positive correlation between the raw counts.

Table 1 Raw Correlations

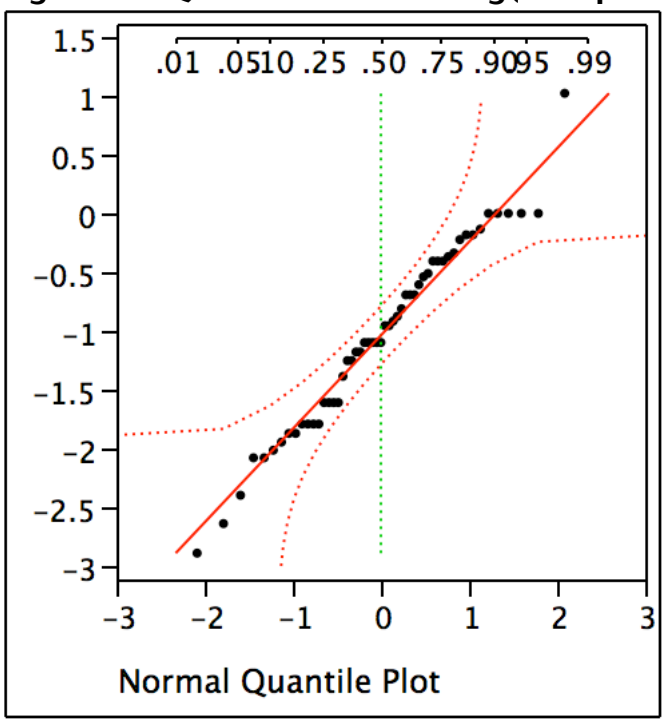
	Creepies	Crawlies	Flies
Creepies	1.0000	-0.3765	0.0861
Crawlies	-0.3765	1.0000	0.0624
Flies	0.0861	0.0624	1.0000

Table 2 Partial Correlations (controlled for Feed)

	Creepies	Crawlies	Flies
Creepies	1.0000	-0.4134	0.0337
Crawlies	-0.4134	1.0000	-0.0794
Flies	0.0337	-0.0794	1.0000

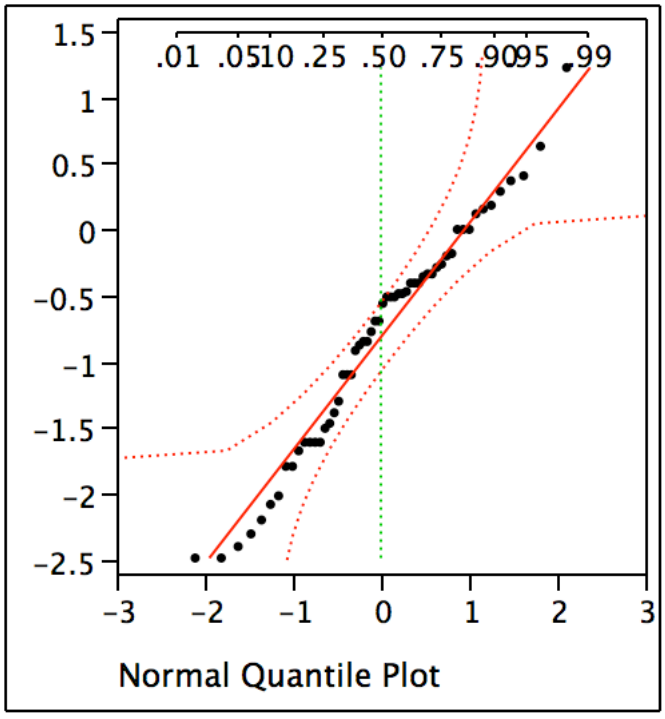
Tables 1 and 2 show the correlation of the counts before and after controlling for the proportion of feeding time. While it is possible to try and apply a logistic normal distribution to the composition of the insects, we have zeros which are not due to limitations in the measurement process and which would have to be taken account of, so it does not make sense to ignore the process information that we have, which may well account for the zeros.

Figure 3a Quantile Plot for Log(Creepies/Flies)



7 missing data

Figure 3b Quantile Plot for Log(Crawlies/Flies)



4 missing data

Interestingly, when we look at the quantile plots of the logratios, in Figures 3a and 3b, which exclude the 11 data points with zeros, the data does not look too different from a normal distribution. However, logistic normal is not the process that generated the data, so it is clear that using the correct process is important. Fortunately, we have a suitable distribution for count data with dependence, which is the Poisson-Log Normal distribution proposed by Aitchison Ho (1989), which is a Poisson distribution mixed with log normal for the means of the Poission distribution. Using this approach makes sense in terms of the underlying process and naturally incorporates the zeros and the discrete nature of the data. Using WinBugs to perform the Bayesian analysis of the Poisson-Log Normal distribution, as suggested by Tunaru (2003), with the extension that the lognormal means are assumed proportional to the feeding time proportions, provides a good fit to the data that supports the graphical evidence in favour of dependence structure amongst the different types of insects. If we only knew the proportion of insects of different types, we can still fit the Poisson-Log Normal distribution together with a compositional process applied by treating the total count as an unobserved integer, which is an easy extension in a Bayesian analysis. In practice, the discrete nature of the composition enables us to generate the underlying counts to a high degree of precision (as the

answers are exact modulo common factors of the counts), so little information is lost by using the relative counts instead of the raw counts.

Bayesian analysis using WinBUGS suggests that there is negligible additional variation in the fly count beyond the Poisson process, while there is strong negative correlation between the additional variation for the creepie and crawlie counts.

3. Modelling the weights

Figure 4 Bivariate Fit of Total Weight By Feed

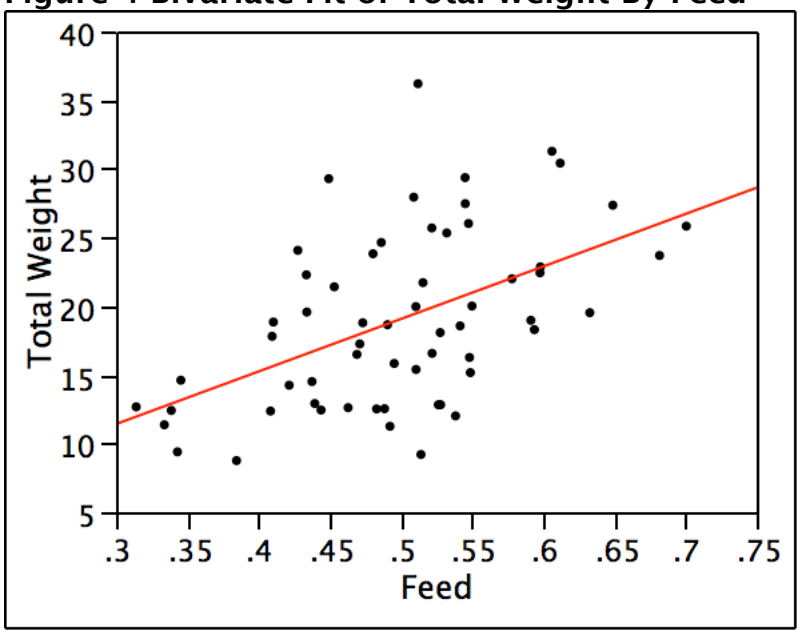


Figure 4 shows a plot of the total weight of insects caught against the proportion of time spent feeding, which also shows a clear linear relationship.

Figure 5a Creepie Wt/insect By Creepies

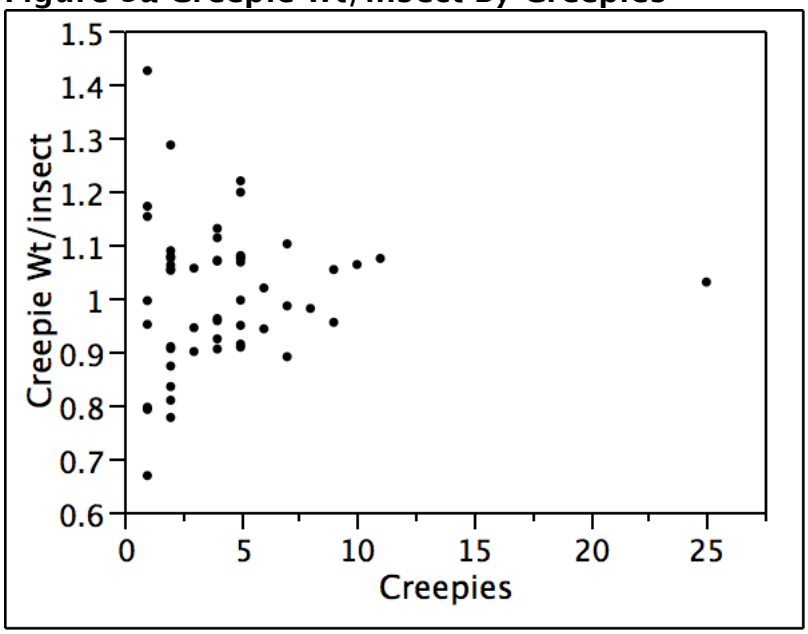


Figure 5b Crawlie Wt/insect By Crawlies

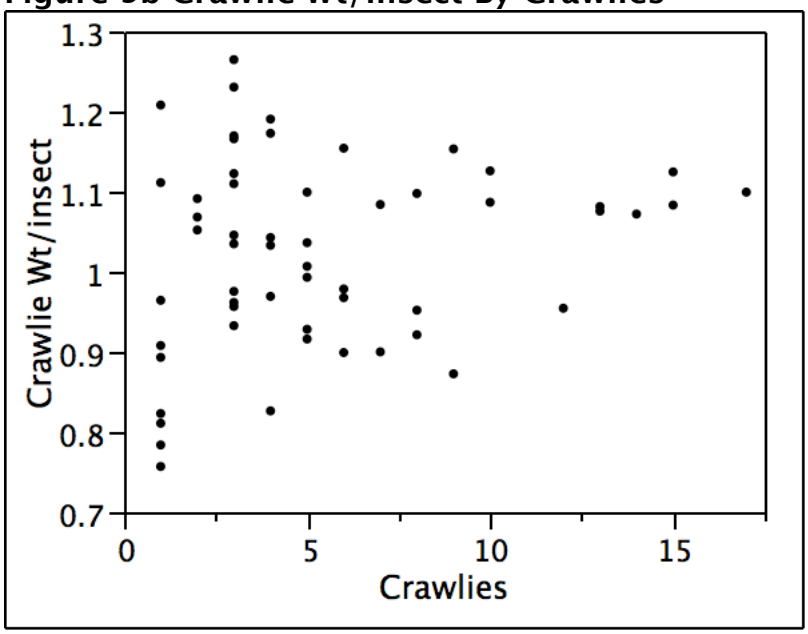
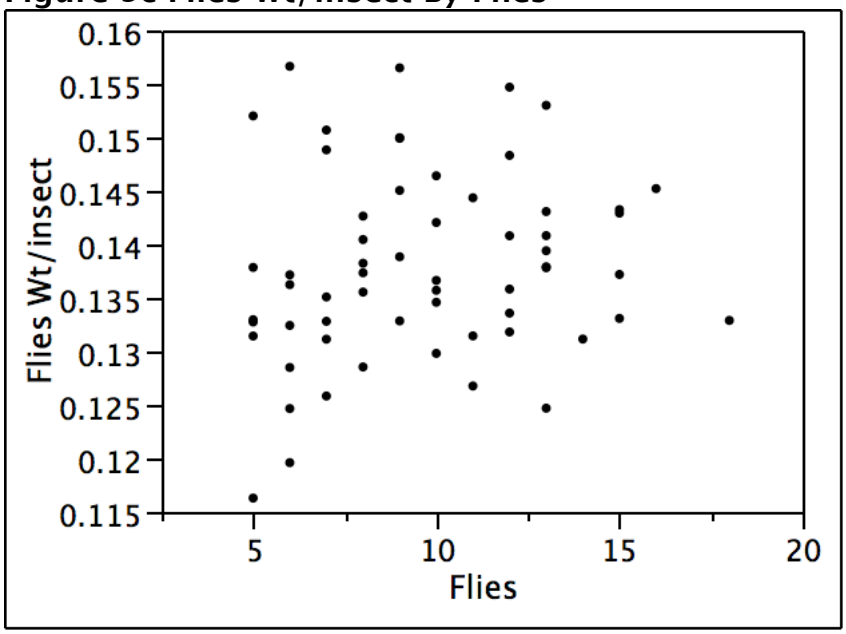
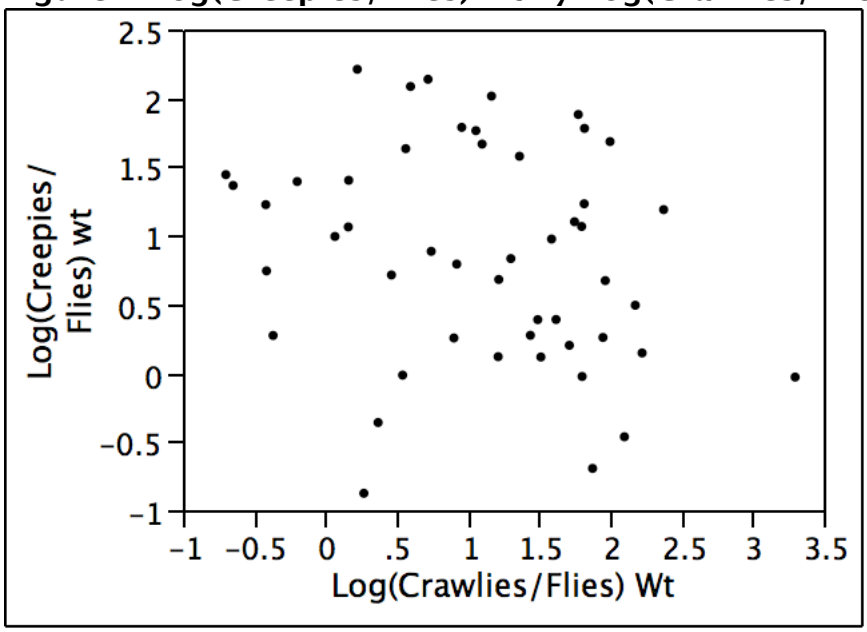


Figure 5c Flies Wt/insect By Flies



However, Figures 5a, 5b and 5c show that the average weight of each insect of each type looks independent of the count for that insect. Careful analysis shows that the weight per insect is fitted best by independent lognormal distributions where the mean is lower for flies than for the creepies and crawlies.

Figure 6 Log(Creepies/Flies) wt By Log(Crawlies/Flies) Wt



11 missing data points

Again, it would be possible to analyse the composition of weights of different insect types (Figure 6) using log ratios, but this ignores the underlying process and would have the continuing problem of zeros

which contain useful information about the process and would require special conditional treatment as suggested by Aitchison and Kay (2003), Bacon-Shone (2003) or Fry et al (2001).

4. Conclusions

This analysis again confirms the value of identifying the correct process that generated the data before applying any statistical analysis. It is clear that, at least in some situations, zeros in compositional data can be correctly handled by modeling an underlying discrete counting process and then perhaps mixing with another process on R^{d+} rather than "adjusting" the zeros to allow the application of the logistic normal process.

5. Ongoing work

The extension of this work that currently interests me is situations with more zeros than would be generated by the Poisson-Log Normal distribution. To model this, I consider the zero inflated Poisson as the underlying process instead of standard Poisson and incorporate this in the manner considered by Aitchison and Kay (2003), Bacon-Shone (2003) and Fry for logistic normal. More also needs to be done in understanding the impact of using logistic normal methods combined with zero replacement as an approximation to the processes considered in this paper.

6. References

Aitchison, J. (2003). Compositional data analysis: where are we and where should we be heading?", CoDaWork 03

Aitchison, J., Ho C.H. (1989). The multivariate Poisson-log normal distribution, *Biometrika*, 76(4), pp. 643-653

Aitchison, J. & Kay, J.W. (2003). Possible Solutions of Some Essential Zero Problems in Compositional Data Analysis, CodaWork 03

Bacon-Shone, J. (2003). Modelling structural zeros in compositional data, CodaWork 03

Fry, J.M, Fry, T.R.L & McLaren, K.R. (2000). Compositional data analysis and zeros in micro data, *Appl. Economics*, 32, pp. 953-959.

Tunaru, R. (2002). Hierarchical Bayesian Models for Multiple Count Data, *Austrian J. of Statistics*, 31, 2&3, pp. 221-229

Appendix Time budget and insect capture data for Goilbirds

Feed	Fight	Perch	Sleep	Creepies	Crawlies	Flies	Creepie Weight	Crawlie Weight	Flies Weight
0.5476	0.0107	0.0113	0.4303	4	13	9	4.281	14.058	1.305
0.5385	0.0253	0.009	0.4271	0	5	7	0.000	5.034	0.881
0.4712	0.0175	0.0211	0.4902	6	2	9	6.112	2.104	1.196
0.483	0.0091	0.0553	0.4526	1	5	6	0.796	4.965	0.817
0.434	0.0031	0.1003	0.4627	7	4	8	7.708	3.305	1.028
0.522	0.0169	0.0321	0.429	5	1	10	4.980	0.964	1.464
0.5939	0.0027	0.0115	0.3919	2	9	8	2.149	7.853	1.099
0.5781	0.0229	0.0222	0.3767	2	8	13	1.818	7.371	1.621
0.4733	0.0047	0.0122	0.5098	5	5	7	6.094	5.497	0.930
0.4863	0.0309	0.0096	0.4732	0	15	8	0.000	16.248	1.141
0.5277	0.022	0.0058	0.4445	4	6	8	3.845	6.925	1.124
0.444	0.0128	0.0044	0.5389	2	4	6	2.155	4.132	0.718
0.5106	0.0076	0.0215	0.4603	4	3	12	3.829	3.329	1.582
0.5264	0.0016	0.0406	0.4313	4	3	6	3.618	2.927	0.748
0.5323	0.0088	0.0262	0.4327	5	5	15	4.540	5.181	2.058
0.4396	0.0119	0.0258	0.5227	4	1	8	4.450	0.908	1.106
0.5981	0.0067	0.0191	0.3761	1	3	18	0.995	3.137	2.392
0.5453	0.0312	0.0121	0.4115	0	15	10	0.000	16.869	1.420
0.3141	0.0063	0.156	0.5236	2	3	7	1.810	3.105	0.918
0.4096	0.0049	0.0227	0.5628	6	1	11	5.653	0.757	1.446
0.463	0.0112	0.0068	0.519	3	1	9	2.833	0.893	1.349
0.3388	0.0073	0.0235	0.6304	1	3	7	1.171	3.509	1.042
0.612	0.0095	0.0107	0.3679	3	10	15	3.167	10.867	2.149
0.5121	0.0063	0.0205	0.4611	25	0	9	25.741	0.000	1.350
0.5489	0.002	0.0149	0.4341	0	6	10	0.000	5.806	1.357
0.4105	0.0011	0.0129	0.5755	9	0	9	8.590	0.000	1.408
0.5107	0.0048	0.0046	0.4798	5	4	6	4.569	4.762	0.771
0.5914	0.0396	0.0116	0.3574	7	1	12	6.896	0.811	1.630
0.55	0.0071	0.005	0.4378	4	3	13	3.694	2.885	1.812
0.5452	0.0171	0.019	0.4186	1	14	14	1.152	15.009	1.836
0.5218	0.0257	0.0477	0.4048	11	3	11	11.812	2.870	1.394
0.4907	0.0046	0.1617	0.3429	2	7	10	1.669	7.588	1.366
0.4085	0.0047	0.0442	0.5425	0	3	9	0.000	3.794	1.250
0.649	0.0143	0.0231	0.3136	9	6	13	9.477	5.395	1.793
0.3846	0.0101	0.0721	0.5333	0	3	5	0.000	3.498	0.581
0.5142	0.0218	0.0323	0.4317	2	3	5	1.745	2.798	0.760
0.4805	0.0504	0.0682	0.4009	2	12	10	2.123	11.454	1.298
0.6062	0.052	0.0137	0.3281	2	13	13	2.572	13.984	1.989

0.4494	0.0251	0.028	0.4975	5	10	13	5.396	11.259	1.860
0.5978	0.0162	0.01	0.3759	1	9	11	0.792	10.380	1.588
0.4533	0.007	0.0128	0.5269	2	3	16	1.618	3.368	2.323
0.5091	0.0075	0.0133	0.4701	10	5	12	10.624	4.641	1.780
0.528	0.0314	0.0428	0.3978	7	0	7	6.231	0.000	0.946
0.4216	0.004	0.029	0.5454	2	4	7	1.553	4.691	1.055
0.5417	0.0066	0.0039	0.4478	2	4	12	2.177	4.171	1.690
0.6328	0.0029	0.0801	0.2842	5	2	12	5.367	2.183	1.856
0.4924	0.0146	0.0418	0.4512	5	1	5	5.988	0.823	0.657
0.6818	0.0126	0.0035	0.3021	5	2	15	5.385	2.136	1.996
0.4337	0.0131	0.0186	0.5346	8	4	10	7.843	3.877	1.346
0.7006	0.0065	0.0167	0.2762	5	8	13	4.743	8.782	1.792
0.4954	0.0032	0.0118	0.4895	2	1	12	2.103	1.111	1.603
0.5156	0.0059	0.0206	0.4579	2	3	15	2.107	3.691	2.144
0.4277	0.0006	0.0367	0.535	1	17	5	0.668	18.689	0.689
0.3431	0.0073	0.0761	0.5734	3	1	5	2.700	0.784	0.665
0.4692	0.0057	0.0068	0.5183	4	0	13	4.519	0.000	1.831
0.4886	0.0578	0.0083	0.4453	0	7	6	0.000	6.300	0.940
0.5483	0.0169	0.0114	0.4234	4	6	6	4.276	5.871	0.795
0.3339	0.0367	0.0348	0.5946	5	1	5	5.335	1.208	0.664
0.3455	0.007	0.098	0.5495	1	8	6	0.951	7.617	0.823
0.4376	0.0279	0.1273	0.4072	1	5	8	1.425	4.580	1.084